Appl Intell (2008) 28: 210–221
DOI 10.1007/s10489-007-0057-z

A case-based approach for characterization and analysis of subgroup patterns

Martin Atzmueller · Frank Puppe

Received: 21 October 2006 / Accepted: 13 April 2007 / Published online: 1 June 2007
© Springer Science+Business Media, LLC 2007

Abstract In general, cases capture knowledge and concrete experiences of specific situations. By exploiting case-based knowledge for characterizing a subgroup pattern, additional information about the subgroup objects can be provided. This paper proposes a case-based approach for characterizing and analyzing subgroup patterns: It presents techniques for retrieving characteristic factors and a set of corresponding cases for the inspection and analysis of a specific subgroup pattern. Then, the set of factors and cases are merged into prototypical cases for presentation to the user. Such an alternative view on the subgroup pattern provides important introspective information on the subgroup objects, that is, the cases covered by the subgroup description: Using drill-down techniques, the user can perform a detailed introspection of a subgroup pattern using prototypical pattern cases. Additionally, these enable a convenient retrieval of interesting (meta-)information associated with the respective subgroup objects.

1 Introduction

Subgroup discovery is a powerful and broadly applicable technique for knowledge discovery in databases. Essentially, subgroup discovery aims to discover interesting subgroups (as subsets of the population) concerning a certain target property of interest, for example, in the subgroup of smokers with a positive family history the risk of coronary heart disease (target property) is significantly higher than in the general population.

The discovered interesting subgroups denote *nuggets* or *chunks* of knowledge. A subgroup is usually easy to interpret depending on a suitable description language, for example, using conjunctive selection expressions. In that sense the subgroup description defining the subgroup objects (cases) stands for itself. Nevertheless, after subgroup patterns have been discovered, methods for subsequent subgroup characterization and analysis can be very useful: Methods for evaluating and browsing a set of subgroup patterns [11, 12], can be applied to obtain further important information about the subgroup objects.

In the context of experience management [10] and case-based reasoning, cases contain specific knowledge of previously experienced, concrete problem situations [1]. Usually, a case consists of a problem description part, a solution part, and additional attached meta-information, for example, a description of the context of the case [9]. In the medical domain, for example, specific cases for patients are collected which do not only include the problem description of the case (given by a set of attribute values) but also additional information, for example, images from x-ray or sonographic examinations. The information contained in the cases can then not only be applied for problem-solving but also for an elaborate inspection of cases in the setting of knowledge discovery methods. In the context of subgroup discovery, presenting a characteristic set of cases can be used for identifying typical problem situations and contexts of a specific subgroup. Such introspective information can then support the user in interpreting the discovered subgroup patterns, by presenting a subgroup in an alternative form.

This paper proposes case-based methods providing characterization and analysis capabilities concerning the subgroup extension, that is, the cases covered by the subgroup

M. Atzmueller (✉) · F. Puppe
Department of Computer Science VI, University of Würzburg,
Am Hubland, 97074 Würzburg, Germany
e-mail: atzmueller@informatik.uni-wuerzburg.de

F. Puppe
e-mail: puppe@informatik.uni-wuerzburg.de

description. First, characteristic factors of the subgroup and their respective strengths are identified. Then, typical and extreme cases characterizing the subgroup are retrieved. The obtained set of factors, the respective cases, and associated meta-information can then provide important additional information: In the medical domain, for example, meta-information such as medical images, the name of the examiner that examined or documented a case, and a typical context of a subgroup pattern can both provide important analytical information and increase the actionability of the pattern. The characteristic factors and cases are summarized by generating a *prototypical pattern case*. This case can then be presented to the user as a representative case for a specific subgroup pattern.

The rest of the paper is organized as follows: Sect. 2 introduces subgroup discovery, subgroup patterns, and characterization techniques. After that, Sect. 3 presents methods for case-based characterization and analysis of subgroup patterns: The first technique obtains a ranked list of the characteristic factors of a subgroup pattern. The second technique enables the analysis and exemplification of subgroup patterns using typical and extreme cases. These approaches are combined into a method for generating *prototypical pattern cases* as a condensed representation of the factors and cases characterizing a given subgroup pattern. Section 4 provides two case-studies in the medical domain. Finally, Sect. 5 concludes the paper with a summary, and points out interesting directions for future work.

2 Subgroup discovery and subgroup patterns

The main application areas of subgroup discovery [15, 22] are exploration and descriptive induction, to obtain an overview of the relations between a (dependent) target variable and a set of (independent) explaining variables. A subgroup pattern is specified by a subgroup description language; its quality is determined by a suitable quality function considering a specific target variable (concept of interest).

The following sections introduce the used knowledge representation, describe subgroup patterns and a method for their statistical characterization.

2.1 General definitions

Let Ω_A be the set of all attributes. For each attribute $a \in \Omega_A$ a range $\mathrm{dom}(a)$ of values is defined; $\mathcal{V}_A$ is assumed to be the (universal) set of attribute values of the form $(a = v)$, $a \in \Omega_A$, $v \in \mathrm{dom}(a)$. Other common names for attribute values are *findings* and *observations*. A *case c* is defined as a tuple

$$c = (\mathcal{V}_c, \mathcal{I}_c),$$

where $\mathcal{V}_c \subseteq \mathcal{V}_A$ is the set of attribute values observed in the case c. The set of attribute values contains the set of *observations* for the given case and also the solution(s) of a case if available, that is, at least one diagnosis in the medical domain. In our context, it is not necessary to explicitly consider the solution part of a case that is usually modeled for case-based reasoning applications. The set $\mathcal{I}_c$ provides additional (meta-)information. The set of all possible cases for a problem domain is denoted by Ω_C. Let $CB \subseteq \Omega_C$ be the case base containing all available cases.

2.2 Subgroup patterns

A subgroup pattern is defined by a subgroup description language. In the single-relational propositional case, a conjunctive subgroup description

$$sd = \{e_1, e_2, \ldots, e_n\},$$

is defined by the conjunction of a set of selectors $e_i = (a_i, V_i)$, that is, selection expressions on domains of attributes, $a_i \in \Omega_A, V_i \subseteq \mathrm{dom}(a_i)$. Consider the subgroup *smokers with a positive family history*, for example: This subgroup, with the target property *coronary heart disease*, is defined by the selectors *smoker* $=$ *yes* and *family history* $= +$. The set Ω_E is defined as the set of all selection expressions and Ω_{sd} as the set of all possible subgroup descriptions

The interestingness of a subgroup pattern can be flexibly formalized by a (user-defined) quality function q

$$q : \Omega_{sd} \to R,$$

that is used in order to evaluate a subgroup description $sd \in \Omega_{sd}$ [15, 22]. Commonly, quality functions include the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size, as criteria for ranking subgroups. Let us consider the exemplary quality function q_S, for binary target variables.

$$q_S = \frac{p - p_0}{\sqrt{p_0 \cdot (1 - p_0)}} \sqrt{n}. \tag{1}$$

For comparing the distribution, the target share in the subgroup (p), estimated by the relative frequency of the target variable in the subgroup, is compared to the default share (p_0) of the target variable in the total population; n denotes the subgroup size. Then, the k best subgroups and/or the subgroups with a quality above a minimum threshold are selected, which usually should have a minimal overlap.

2.3 Statistical characterization of subgroup patterns

Subgroups can always be characterized by the factors used to describe them, that is, by the selectors contained in the subgroup description. However, besides these *principal factors* there are certain *supporting factors* that can also be applied in order to characterize a subgroup, c.f., [11]: The supporting factors are given by attribute values $supp \subseteq \mathcal{V}_A$ contained in the subgroup cases that are identified using basic statistical analysis. The value distributions of their corresponding (supporting) attributes differ significantly comparing the subgroup and the total population with respect to the concept of interest. Supporting factors for the subgroup *smokers with a positive family history* are given by the factors $hypertension = yes$ or $overweight = yes$, with respect to the target property *coronary heart disease*. Thus, given a binary target variable, a supporting attribute a of a subgroup s is defined as an attribute with a significantly different distribution comparing the true positive (target class) cases contained in the subgroup s and all the negative (non-target) cases contained in the total population.

Then, an attribute value $(a = v)$ corresponding to the selector $e = (a, \{v\})$ of a supporting attribute a is characteristic for the subgroup, that is, it is a supporting factor, if it is positively associated with the true positive (target class) cases contained in the subgroup compared to all the negative cases. The statistical significance of an attribute and an attribute value is tested using the standard χ^2-test for independence with a 0.05 significance level (that is, with a confidence level of 95%) and the correlation- or ϕ-coefficient for binary variables, respectively.

The principal factors can be regarded as *strong* factors, while the supporting factors can be regarded as a kind of *weak* factors: The principal factors are observed in all cases of a subgroup while the supporting factors are only observed in some cases. Nevertheless, the supporting factors can provide important additional information with respect to the target cases contained in the subgroup. As discussed by Gamberger et al. [11] the supporting factors that characterize the subgroup can also be very helpful for the user: Given the principal factors the supporting factors can provide additional evidence with respect to the target concept. In this way, observing the supporting factors can facilitate an easier recognition of target cases [16]: If a case is assigned to a subgroup based on the principal factors, then observing a supporting factor provides for some evidence that the case is potentially positive with respect to the concept of interest. Thus, the supporting factors are used to point at specific characteristics of the target space covered by the subgroup. Then, the set of the *characteristic factors* is defined as the union of the principal and supporting factors.

The principal factors are contained in all cases of a subgroup, whereas the supporting factors do not occur in all cases but may occur in many cases of the subgroup. Then, their individual strength in confirming the concept of interest, that is, their *relative importance* can be scored, as described in Sect. 3.1.

3 Case-based subgroup characterization and analysis

This section describes the methods of the proposed approach for case-based subgroup characterization and analysis: Given a specific subgroup pattern, first a set of characteristic factors (selectors) for the subgroup is obtained. Next, these factors are ranked in order to obtain a set of exemplifying (real) cases for the given factors. After that, a *prototypical pattern case* is created capturing the characteristic factors of the subgroup pattern, the set of characteristic and exemplifying cases, and a set of relevant additional factors.

The approach for case-based characterization and analysis of subgroup patterns consists of the following steps shown in Fig. 1:

1. Given a subgroup pattern s, the characteristic factors given by a set of selectors $F \subseteq \Omega_E$ are extracted.
2. Next, the obtained characteristic subgroup factors F are scored: For each selector $e \in F$ its respective (confirmation) strength is obtained with respect to the target concept. The assigned scores are then mapped to weights denoting the importance of the respective factors.
3. After that, a case-based retrieval method is applied: Concerning the cases contained in the subgroup either typical or extreme cases with a high coverage of the characteristic factors F are retrieved—as exemplifying cases for the

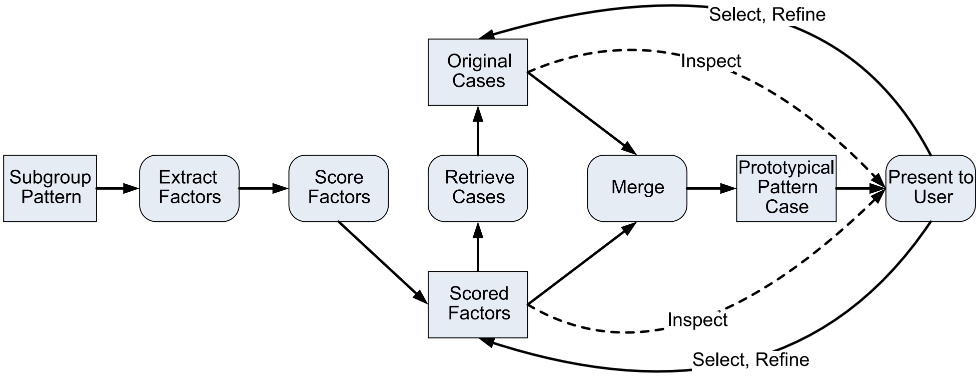


Fig. 1 Process model: Case-based characterization and analysis

subgroup pattern. In the retrieval method the factors can be weighted according to their relative importance, that is, according to the assigned weights, depending on the requirements of the user.

4. Finally, the scored characteristic factors and the retrieved cases are merged into a virtual prototypical pattern case. This case is then presented to the user in order to facilitate an easier interpretation.

The proposed process shown in Fig. 1 is incremental and can include user feedback. For this purpose, there are several 'shortcuts' in order to implement user-guided interaction strategies: The user can optionally inspect, select and refine the set of characteristic factors that are considered in the scoring and the retrieval step. Furthermore, the user can also optionally inspect a preview of the retrieved cases before the prototypical pattern case is generated, and can refine or extend this set as well, if needed.

A prototypical pattern case summarizes both the set of the (scored) characteristic factors, the set of the relevant (retrieved) *subcases*, and other selected factors obtained from the set of subcases. The prototypical case representation serves several purposes:

- The user usually first considers the different factors (with assigned confirmation strengths) of the prototypical pattern case. These factors characterizing the subgroup pattern are also reflected by the collection of subcases. The prototypical pattern case can thus be regarded as an extended representative case: It can either contain a summary of the typical problem setting of the subgroup, or a range of the extreme settings of the subgroup pattern.
- Furthermore, the set of the typical or extreme cases of the subgroup can be inspected in detail by the user: The prototypical pattern case also contains a mapping from each subcase to the set of the most similar subcases in order to identify clusters representing related situations in a specific context.
- Each factor contained in the prototypical pattern case is also linked to the originating (real) cases contained in the case base. Then, the different real world situations in which the factors occurs can be inspected by the user. Furthermore, these links provide the opportunity to locate other relevant meta-information.

The following section first shows show how to rank the characteristic factors: The importance of each individual factor is estimated with respect to the target concept in the subgroup. After that, the next section describes the techniques for characterizing and exemplifying subgroup patterns in terms of cases, utilizing methods from case-based reasoning. Then, the approach for generating prototypical pattern cases is presented.

3.1 Scoring subgroup factors

After the set of characteristic factors has been determined, it can already be used for characterizing a subgroup pattern. However, by analyzing these factors further, the (confirmation) strength of each of the characteristic factors F of a given subgroup can be estimated with respect to the target concept. The confirmation strengths basically correspond to the relative importance of the respective factors, and can therefore also be integrated as weights in the case-based retrieval method discussed below. To facilitate an easier interpretation by the user, a restricted set of symbolic categories is utilized. Essentially the individual strength of a factor $e \in F$ is measured with respect to the evidence it provides for the target concept in the subgroup. It is easy to see that the principal factors will always obtain the strongest confirmation category, while often weaker categories will be assigned to the supporting factors.

For rating the subgroup factors concerning their confirmation strengths, two populations are compared: The true positives contained in the subgroup and the false positives of the total population. In this way, it can be estimated how significantly a selector discriminates between the cases containing the target concept in the subgroup, and all the remaining non-target class cases. In the medical domain, for example, a typical analysis aims at identifying factors that are characteristic for a subpopulation of all the patients with a certain disease compared to all the healthy patients.

Scoring the characteristic subgroup factors relies on an adaptation of a method presented in [5]: Given a subgroup, a characteristic factor $e \in F$, and the target concept t, a 2×2 contingency table is constructed—similar to the technique for identifying the supporting factors. With the given target concept t and the selector e corresponding to the given factor, two binary variables T measuring the target class cases contained in the subgroup, and E identifying the cases containing the specific selector e are constructed. The analysis is limited to the cases $\mathcal{C} \subseteq CB$ from the population in which attribute a is not unknown.

Considering a case c, the value *true* is assigned to the variable T, if c is contained in the subgroup s and if the target variable t occurs in c, and false otherwise; the value *true* is assigned to a variable E, if the selector e occurs in a case, otherwise E is false. The four-fold table is filled as shown below:

	$T = true$	$T = false$
$E = true$	a	b
$E = false$	c	d

The frequency counts denoted in the table are defined as follows:

$a = N(T = true \wedge E = true)$,

$b = N(T = false \wedge E = true)$,

$c = N(T = \mathit{true} \wedge E = \mathit{false})$,

$d = N(T = \mathit{false} \wedge E = \mathit{false})$,

where $N(\mathit{cond})$ is the number of times the condition *cond* is true for cases $c \in \mathcal{C}$.

Next, the distribution of the factor of the true positives in the subgroup, that is, the target class cases, is compared to all negative cases. By definition, this association is always significant concerning the characteristic factors. Then a score $s \in [0; 1]$ is computed according to the strength of the association, and mapped to a symbolic category which is assigned to the respective selector. The general approach is summarized in the following algorithm, and the individual steps are explained in detail below.

1. **Require:** Subgroup s, target variable t, a selector e.
2. Construct binary variables T, E that identify the target class cases of the subgroup s compared to all negatives of *CB*, and the cases where e occurs.
3. Compute the quasi-probabilistic score *qps*, $\mathit{qps} = \mathit{prec}(r) \cdot (1 - \mathit{FAR}(r))$ with respect to the pseudo-rule $r : e \rightarrow t$.
(see below for definitions of *prec* (precision) and *FAR* (false alarm rate)).
4. Map the *qps*-score to a symbolic category s using a conversion table.
5. Label the selector e with the obtained symbolic category.

For determining the respective scores two well-known measures are utilized: *precision* and the *false alarm rate (FAR)*, which is also known as the *false positive rate*, or 1—specificity. The precision of a selector e is defined as

$$\mathit{prec}(e) = \frac{TP}{TP + FP}, \tag{2}$$

whereas the false alarm rate *FAR* for a selector e is defined as

$$\mathit{FAR}(e) = \frac{FP}{FP + TN}. \tag{3}$$

The symbols *TP*, *TN*, *FP* denote the number of *true positives*, *true negatives*, and *false positives*, respectively. These can easily be extracted from the contingency table. For a positive dependency between selector e and a target variable t, the parameters are defined as $TP = a$, $TN = d$ and $FP = b$.

Since the *quasi probabilistic score (qps)* for a selector e combines the precision and the false alarm rate, $\mathit{qps}(e) = \mathit{prec}(e) \cdot (1 - \mathit{FAR}(e))$, the qps-measure achieves a trade-off between the accuracy of the selector to predict a target variable measured against all predictions weighted by the proportion of false predictions. It is worth noting, that often the *true positive rate (TPR)*—which is also known as *recall/sensitivity*—is used in combination with the FAR as a measure of accuracy, for example, combined by the *F-Measure* [21, Chap. 5]). However, this is mostly applicable to standard rules, which usually contain more complex rule conditions than using single selectors. For descriptive purposes, the scoring relations are assessed independently but can support each other in providing evidence for the target variable. Thus, their accuracy needs to be assessed on localized regions of the target space. In this case, precision is more suggestive, since it does not take the complete target space into account, but it measures only the accuracy of the localized factor due to the independence assumption.

Next, there are two options for utilizing the score: First, the obtained score can be mapped to a symbolic confirmation category $sc \in \{+, ++, +++\}$ that specifies confirming symbolic categories in ascending order. Using a suitable conversion table, for example, given by the mapping of the intervals $[0, 0.5)$, $[0.5, 0.9)$, $[0.9, 1.0]$ to the respective categories $\{+, ++, +++\}$. The loss of information is neglected in order to increase the understandability of the learned relations.

The symbolic category *sc* expresses the strength or the relative importance of a given selector e. For each factor (selector) $e \in F$ a scoring selector $e' = (e, sc)$ is constructed assigning the respective confirmation category *sc*. Then, the scored selectors can be presented to the user for an intuitive overview of the important factors and their corresponding strength for confirming the target concept of the subgroup. Second, the obtained scores can be utilized for the case-based retrieval method described below: Since the confirmation categories denote the strength of the association between an individual factor and the target concept of the subgroup, the individual categories can be directly mapped to weights denoting the relative importance of the factors. The weights can then be applied in the retrieval method when estimating the similarity of cases.

3.2 Identifying exemplary cases for subgroup patterns

As a first step for analyzing a specific subgroup pattern a set of exemplary cases of the pattern is retrieved: This step aims at utilizing the implicit experiences contained in the cases of the case base as explaining examples. Given a set of characteristic factors F of the subgroup or a user-selected subset of these, either typical or extreme cases with a high coverage of the set of factors F can be retrieved. By inspecting these sets of cases 'as is' the user is already able to obtain a view on the general 'problem setting' of the subgroup. The next step combines these cases and the factors into a prototypical case as an intuitive alternative form. The next section describes how the factors and the cases are merged into a prototypical pattern case as a condensed representation.

For exemplifying a subgroup pattern, a naive solution retrieves all the target class cases contained in the subgroup. However, this approach suffers from two shortcomings: First, the set of cases can be quite large for a comprehensive overview. An extensive analysis and manual inspection of all the cases is then intractable. Furthermore, a subset of F is not accounted for very precisely, that is, the supporting factors: The target class cases contained in the subgroup are determined by the set of principal factors contained in the subgroup description, and the target concept only. In contrast, the supporting factors might cover quite a diverse set of cases, since they are not included in all of the cases. In order to include the relative importance of the individual factors during the case-retrieval step, the individual strengths of the factors can be taken into account utilizing the learned weights. This can also be extended using background knowledge, for example, by utilizing partial similarities between attribute values, if available.

Case retrieval The case retrieval step aims at retrieving a set of (target-class) cases contained in the subgroup that have a high coverage with the set $F \subseteq \Omega_E$ containing the characteristic factors (selectors). Then, there are two options to characterize the set F:

- First, *typical* cases can be retrieved that are most similar to F while the individual cases can also be very similar to each other. These cases can then be used to exemplify the most common factors contained in F.
- Second, *extreme* cases can be retrieved, that is, cases that are very similar to F but not to each other. This set of diverse cases is discriminative and can be used in order to obtain a comprehensive view on the setting of extreme factor combinations concerning the set F.

For the retrieval step techniques adapted from case-based reasoning methods [1] are applied: Given a query case q, k similar cases $\{c_1, \ldots, c_k\}, c_i \in CB$ are retrieved. The attribute values contained in the query case are commonly called the *problem description*. Considering a *virtual* query case q, its problem description is defined as the set of characteristic factors F_i obtained from a given subgroup s_i. Optionally, the user can modify and tune F_i interactively to fit the analysis requirements: A subset F' of the factors F_i can be selected, for example, the most interesting factors.

For assessing the similarity of a virtual query case q and a retrieved case c, the *matching features* similarity function $\mathrm{sim}(q, c)$ given in Equation 4 can be used, if there is no pair of selectors in F' containing the same attribute:

$$\mathrm{sim}(q, c) = \frac{|\{e \in F' : e = \pi_e(c)\}|}{|F'|}. \tag{4}$$

The considered factors include the factors $F' \subseteq F_i$ contained in the query case q; the function $\pi_e(c)$ returns the selector $e = (a, \{v\})$ corresponding to the value of attribute a of the original case c.

Alternatively, a weighted similarity measure given in (5) can be applied by taking the learned weights (given by the scores) of the factors into account. This is also necessary, if F' contains a pair of selectors for the same attribute. Therefore, for the similarity measure only the set of attributes contained in the query case q are considered, that is, the attributes $\Omega'_A = \{a \in \Omega_A \mid \exists e \in F', e = (a, V_a)\}$. Then, $\pi_a(c)$ returns the set of selectors corresponding to the attribute a for a virtual query case c, and a set containing one single-valued selector for an original case c.

$$\mathrm{sim}(q, c) = \frac{\sum_{a \in \Omega'_A} w(\pi_a(q), \pi_a(c)) \cdot \mathrm{match}(\pi_a(q), \pi_a(c))}{\sum_{a \in \Omega'_A} w(\pi_a(q), \pi_a(c))}. \tag{5}$$

The matching function $\mathrm{match}(\pi_a(q), \pi_a(c))$ returns 1 if $\pi_a(c) \subseteq \pi_a(q)$, and 0 otherwise (A possible extension could consider similarities of selectors providing values between 0 and 1). $w(\pi_a(q), \pi_a(c))$ returns the weight of the selector corresponding to the attribute value contained in the case c if $\pi_a(c) \subseteq \pi_a(q)$, and the maximum weight of the selectors contained in $\pi_a(q)$ otherwise.

The diversity of a set of retrieved original cases $\mathcal{RC} = \{c_i\}_k$ of size k is computed according to the measure $\mathrm{diversity}(\mathcal{RC})$, defined as follows:

$$\mathrm{diversity}(\mathcal{RC}) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (1 - \mathrm{sim}(c_i, c_j))}{k \cdot \frac{(k-1)}{2}}, \tag{6}$$

where the similarity of two original cases is estimated analogously as described above. To retrieve the set of the most extreme cases with respect to a subgroup pattern, techniques that obtain a set of most similar but diverse cases regarding to the query case are applied. There are several methods to retrieve a set of diverse cases as described, for example, in [17]. The presented approach applies the *Bounded Greedy (BG)* algorithm introduced by Smyth and McClave [20]: BG starts with a retrieval set initially containing the most similar case to the query case. In each iteration of the algorithm the case in the set of the $2k$ most similar cases is selected which maximizes both the product of its similarity to the query case and its relative diversity with respect to the cases that have been selected for the retrieval set so far.

The relative diversity $\mathrm{relDiversity}(c, RC)$ of a case c with respect to the retrieval set $\mathcal{RC} = \{c_i\}_m$ of size m is defined as

$$\mathrm{relDiversity}(c, RC) = \frac{\sum_{i=1}^{m} 1 - \mathrm{sim}(c, c_i)}{m}. \tag{7}$$

BG stops if the retrieval set reaches its pre-specified size k. To obtain a smaller number of diverse (extreme) cases, optionally the smallest subset $R' \subseteq R$ can be selected, for which the coverage between the problem description of a query case q and the union of the problem descriptions contained in R' is maximized.

The retrieved set of typical (or extreme) cases can be seen as a set of explaining examples for the given set of factors characterizing a specific subgroup. Thus, a subgroup can be inspected in a different view by considering specific exemplary cases. By presenting typical or extreme cases the user obtains a detailed and intuitive impression about the objects (cases) contained in the subgroup. The next section describes how to merge the retrieved cases into a prototypical pattern case for a convenient presentation to the user.

3.3 Generating prototypical pattern cases

This section shows how to generate prototypical cases in order to create a representative of the retrieved typical or extreme cases of a subgroup pattern and its characteristic factors. A *prototypical pattern case*

$$cp = (E_{cp}, C_{cp}, \sigma_{cp}, \delta_{cp})$$

consists of a set of subcases $C_{cp} \subseteq CB$ of a given case base CB, a set of selectors $E_{cp} \subseteq \Omega_E$ containing the characteristic factors and additional relevant selectors, a mapping function relating a selector of the constructed prototypical case to its set of (originating) subcases

$$\sigma_{cp} : E_{cp} \rightarrow 2^{C_{cp}},$$

and a case selection function

$$\delta_{cp} : C_{cp} \times \mathbb{N} \rightarrow 2^{C_{cp}}.$$

The selection function δ_{cp} returns a set of the most similar l subcases for a specific subcase of the prototypical pattern case cp, for which $l \in \mathbb{N}, l \leq k = |C_{cp}|$.

It is easy to see that all the selectors included in the generated virtual query case can be transferred to the prototypical pattern case: These factors are given by the set of characteristic (principal and supporting) factors (or a user-selected subset of these). For the remaining attributes contained in the subcases that are not included in the set of characteristic attributes a discriminative set of *additional selectors* needs to be selected: Based on the respective attribute values, the prototypical case is padded with a set of selectors characterizing the subcases. However, when combining the respective attribute values, conflicts can arise if two cases contain different values for the same attribute. Therefore, a conflict resolution strategy for competing attribute values of a specific attribute is applied: Either a majority vote is drawn or background knowledge can be applied, if available:

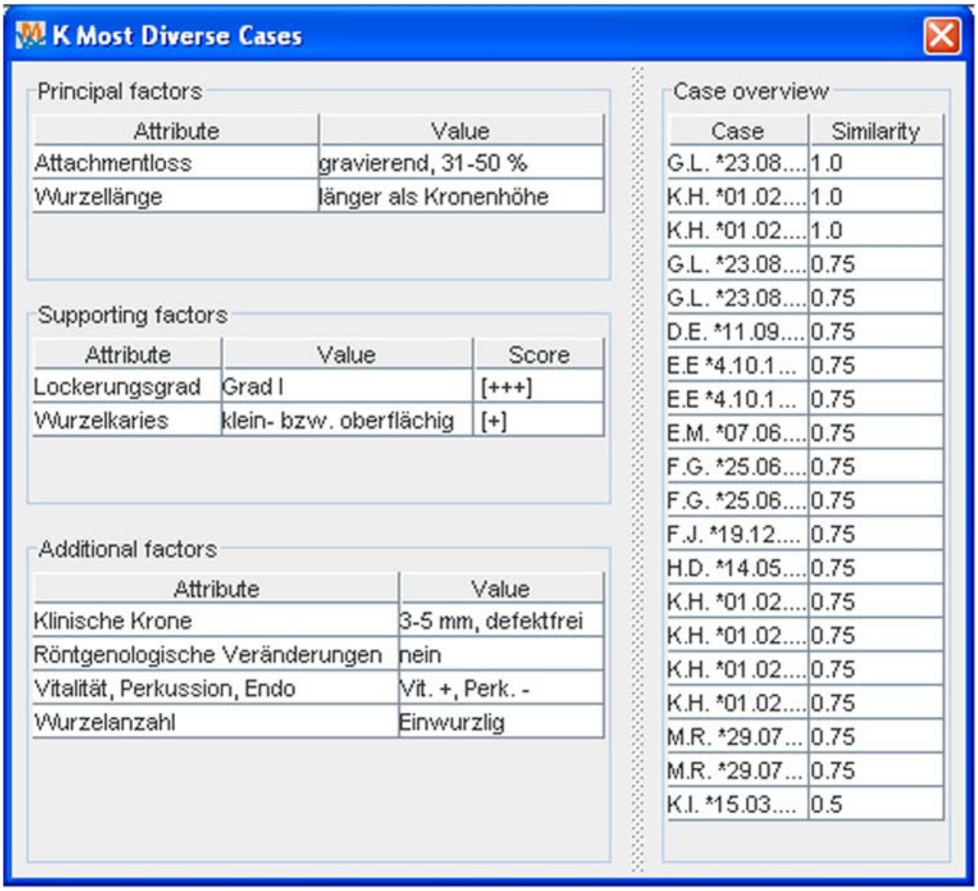
1. Generally, the most frequently occurring value v from the set of the respective attribute values V contained in the subcases is selected, that is,

   $$v = \arg\max_{v_i}(\text{freq}\{v_i \in V\}).$$

   In the case of ties, the value that is most positively associated with the target concept is selected utilizing the technique described in Sect. 3.1.
2. Alternatively, background knowledge can be applied, if available: One element of applicable background knowledge is given by *abnormality knowledge* [6] which is quite common in some domains, for example, in the medical domain. Abnormality knowledge specifies which attribute values represent a normal or an abnormal state of their corresponding attribute, e.g. the value *pain* = *none* is normal, whereas the value *pain* = *high* is abnormal for a certain attribute/symptom.
   If abnormalities are defined, then the value with the highest abnormality is selected. This approach is motivated by the heuristic that often the abnormal values are more interesting for the user, for example, in the medical domain. Considering two patients with two (different) diseases, for example, then it seems to be reasonable that the more severe attribute value (finding) will be selected. Then, *pain* = *high* from one diagnosis is chosen rather than *pain* = *none* from another one. This is especially helpful when considering a set of extreme cases characterizing the subgroup, since the abnormal values indicate extreme conditions.

The set of selectors of a generated prototypical pattern case is then given by the set of principal factors and supporting factors of a given subgroup pattern, and by additional factors contained in a set of exemplifying cases. These are added in order to fill the prototypical pattern case with representative values contained in the subcases, and provide a broader context for the prototypical pattern case. The mapping function σ_{cp} of a prototypical pattern case *cp* is modeled by creating a link from each selector of the case *cp* to the set of the original subcases containing the value, when merging the set of selectors.

Both the selection and the mapping function enable a 'drill-down' approach when further analyzing a set of factors or a set of cases: The user can easily inspect a related set of subcases, and can also inspect each originating case for a specific attribute value. However, if the user wants to 'drill-down' on the sets of the selectors, retrieving a set of typical cases may result in a set of subcases that do not contain some of the characteristic factors. Especially if the weight of a certain factor is quite low, then the typical cases might be focused on the factors with a higher weight. Retrieving a set of extreme cases can then be an option in order to cover the whole set of the characteristic factors.



Principal factors

Attribute	Value
Attachmentloss	gravierend, 31-50 %
Wurzellänge	länger als Kronenhöhe

Supporting factors

Attribute	Value	Score
Lockerungsgrad	Grad I	[+++]
Wurzelkaries	klein- bzw. oberflächig	[+]

Additional factors

Attribute	Value
Klinische Krone	3-5 mm, defektfrei
Röntgenologische Veränderungen	nein
Vitalität, Perkussion, Endo	Vit. +, Perk. -
Wurzelanzahl	Einwurzlig

Case overview

Case	Similarity
G.L. *23.08....	1.0
K.H. *01.02....	1.0
K.H. *01.02....	1.0
G.L. *23.08....	0.75
G.L. *23.08....	0.75
D.E. *11.09....	0.75
E.E *4.10.1 ...	0.75
E.E *4.10.1 ...	0.75
E.M. *07.06....	0.75
F.G. *25.06....	0.75
F.G. *25.06....	0.75
F.J. *19.12....	0.75
H.D. *14.05....	0.75
K.H. *01.02....	0.75
K.H. *01.02....	0.75
K.H. *01.02....	0.75
K.H. *01.02....	0.75
M.R. *29.07...	0.75
M.R. *29.07...	0.75
K.I. *15.03....	0.5

Fig. 2 An exemplary screenshot (in German) taken from a knowledge-refinement application in the domain of dental medicine: It depicts the prototypical pattern case for the subgroup *attachmentloss* = *strong* (*Attachmentloss* = *gravierend 31–50%*) *AND root length* = *longer than crown length* (*Wurzellänge* = *länger als Kronenhöhe*) (with respect to the target concept *incorrect recommendation for tooth extraction*). The left pane contains the principal factors, the supporting factors (*toothlax=minor* (*Lockerungsgrad* = *Grad I*, *root caries* = *minor* (*Wurzelkaries* = *klein bzw. oberflächig*)) and their associated scores, and the additional factors of the prototypical pattern case. The right pane shows the retrieved subcases, that is, the 20 most diverse cases for the exemplary subgroup pattern

Figure 2 shows an exemplary screenshot of a prototypical pattern case for a knowledge refinement setting in the domain of dental medicine: The figure depicts the subgroup *attachmentloss* = *strong* (*Attachmentloss* = *gravierend 31–50%*) *AND root length* = *longer than crown length* (*Wurzellänge* = *länger als Kronenhöhe*), and shows the principal factors, the supporting factors and their strengths, other additional factors, and the set of subcases of the generated prototypical pattern case. In this example, inspecting the cases can then help in performing the knowledge refinements, as discussed in Sect. 4.1.

3.4 Discussion

Characterizing subgroup patterns by a set of supporting factors has been proposed by Gamberger et al. [11, 12]. The methods for obtaining the supporting factors and for ranking these can be regarded as being related to correlation-based methods for relevance analysis of attributes and attribute values. However, the supporting factors focus on descriptive aspects of a subgroup pattern, in comparison to approaches for estimating the importance or the relevance of attribute values [13] and for learning weights of attributes [2] in a case-based reasoning context. Thus, the importance of the attributes is estimated with respect to a pattern and a specific target concept, and not concerning the class only: The supporting factors can characterize the subgroup in a different way, orthogonal to the subgroup description.

In contrast to only obtaining the supporting factors (and thus also a subset of the characteristic factors), these are further ranked in order to obtain their confirmation strength for the target concept. The confirmation strengths are given by symbolic categories in order to enable an intuitive interpretation for the user. Furthermore, these can be directly mapped to weights (denoting their relative importance) for the similarity measure used in the case-based retrieval method.

Using prototypical cases has been introduced early in the field of case-based reasoning [8], and is often applied in medical domains [19]. In contrast to the existing approaches, the presented approach does not just aim at summarizing or describing a set of cases. Instead, it focuses on characterizing subgroup patterns: First, characteristic factors are obtained using statistical analysis. Using these sets of exemplifying cases are retrieved. After that, both are combined into a prototypical pattern case, a process for which background knowledge can be included, if available. A prototypical pattern case $cp = (E_{cp}, C_{cp}, \sigma_{cp}, \delta_{cp})$ provides a comprehensive and condensed alternative representation of a subgroup pattern in the form of a single case: The set E_{cp} contains the characteristic factors and a summary of the remaining factors contained in the subcases. The links between the factors and the set of subcases C_{cp} provide an intuitive approach for inspecting the important factors in their specific context, that is, embedded in their originating cases. Furthermore, associated meta-information can be conveniently identified. Finally, the set C_{cp} and the selection function δ_{cp} facilitate an easy inspection and traversal of the neighborhood of exemplifying cases with respect to the given subgroup pattern; relevant meta-information can also be located quite easily.

In this way, a prototypical pattern case provides for a concise, easy to interpret, and transparent representation for analyzing, summarizing and characterizing a specific subgroup pattern. Usually, when inspecting and analyzing a mined subgroup pattern the user will select a small number of k, $k = 10$ or $k = 20$, for example, in order to obtain a summary of the characteristic (sub-)cases that is still easy to comprehend. The number of k can also be varied interactively for a detailed analysis. Furthermore, browsing the set of typical and extreme cases, and a comparison of these can also provide important information.

The main function of a prototypical pattern case is its capability of summarizing the characteristic factors of a subgroup pattern, while a detailed analysis can still be performed by inspecting the contained subcases: This is also necessary, since a prototypical case does usually not correspond to an original (real-world) case, because its problem description is composed of specific factors extracted from a set of cases. However, starting with the problem description of a prototypical pattern case, the user can always apply 'drill-down' techniques on the original cases for inspecting these in more detail, and to obtain additional (meta-)information.

4 Application—case studies

The presented approach has already been successfully applied in the medical domain. The following sections sketch two case studies for the application of the presented method: The first case study was performed with respect to a knowledge refinement setting applying subgroup discovery techniques in the domain of dental medicine. The goal was to improve a given knowledge-base by analyzing subgroup patterns denoting patterns with a high share of erroneous diagnoses. Then, the knowledge base could be extended by modifying and adding new relations as needed. A further study describes an application in the domain of sonography utilizing cases from the SonoConsult system: Subgroup discovery is applied as a technique for knowledge discovery and for quality control. Then, the discovered subgroup patterns can be conveniently analyzed using the case-based techniques.

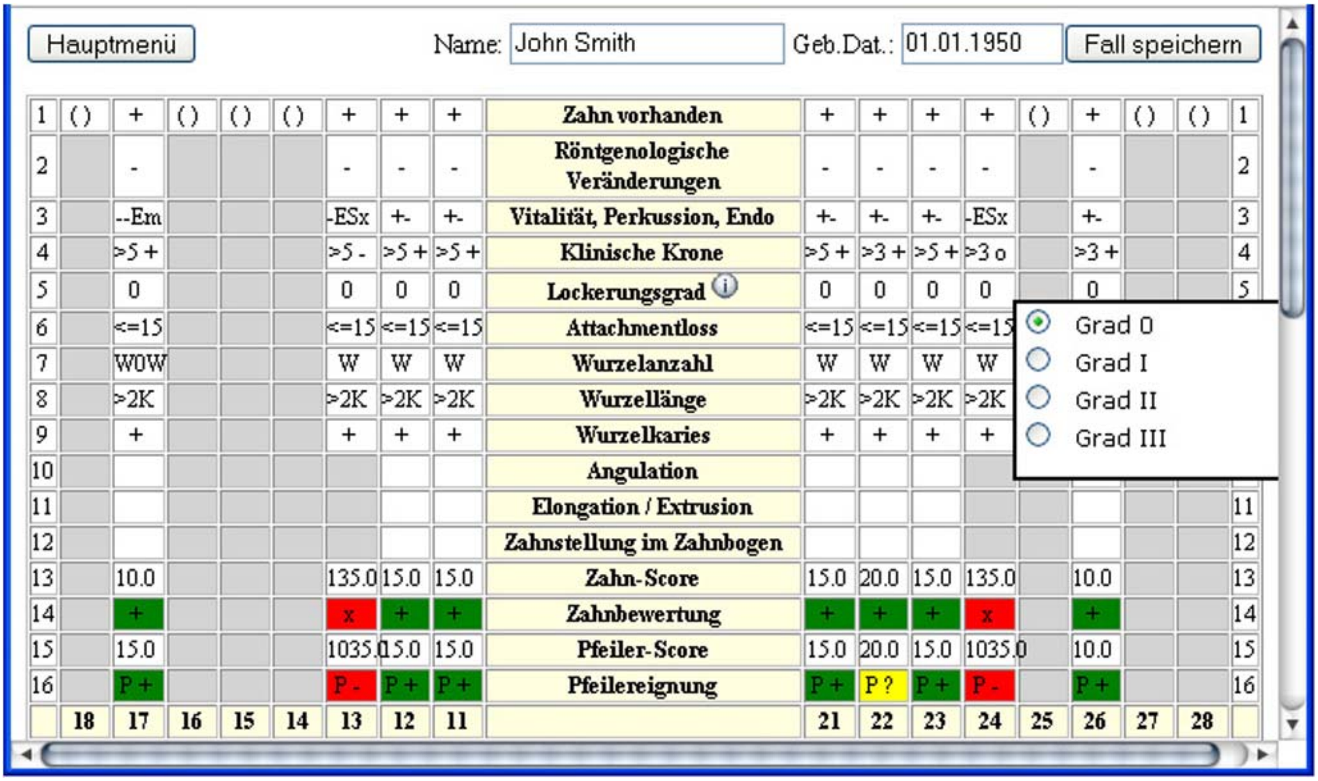
4.1 Analyzing subgroup patterns in the context of interactive knowledge refinement

The first case study concerns the domain of dental medicine were subgroup mining was applied for interactive knowledge refinement of a knowledge-based system. The case study was performed in the domain of dental medicine implemented with a consultation and documentation system for dental findings regarding any kind of prosthetic appliance. The system has been developed in cooperation with the department of prosthodontics at the Würzburg University Hospital.

The system aims to support the decision whether to extract a tooth or not using the documented findings: The cases always contain the standard anamnestic findings and additional findings from x-ray examinations, for example, abnormal x-ray findings ('Röntgenontologische Veränderungen'), grade of tooth lax ('Lockerungsgrad'), endodontic state ('Vitalität, Perkussion, Endo') , root quantity ('Wurzelanzahl'), root length ('Wurzellänge'), crown length ('Klinische Krone'), level of attachment loss ('Attachmentloss'), root caries ('Wurzelkaries'), tooth angulation and elongation/extrusion ('Elongation/Extrusion'), see Fig. 3. For decision support the system derives two distinct diagnosis *EX* and *IN* that either indicate the teeth that could be conserved (IN) or should be extracted (EX). Figure 3 shows a screenshot of the user interface of the system.

The method for knowledge-refinement using subgroup mining methods was successfully applied, improving the correctness of the knowledge base that initially was in an earlier state: The knowledge base was improved significantly by adding and modifying relations that were identified using a subgroup mining approach, cf., [3, 4]. Subgroup mining was applied for pointing at certain subgroups corresponding to 'hot spots' of the knowledge base, that is, specific factor combinations for which the error rate of the system increased significantly. These subgroups were then analyzed by the domain specialists in order to perform refinement operators on the knowledge base, for example, modifying relations or adding new ones. Table 1 shows examples of the discovered subgroups with respect to the incorrect derivation of the diagnosis 'EX' indicating tooth extraction: Subgroup #1 is an example for a simple modification, for which only the (positive) derivation strength needed to be adapted. For subgroup #2 the condition *root length* = *longer than crown length* counts as negative for extraction, and relativizes the factor *tooth lax* = *medium* which is positive for extraction, therefore the knowledge base was modified in order to include this subgroup as an exception.

However, the experiences obtained throughout the earlier parts of the case study motivated the development of further methods for subgroup characterization and analysis: Often small 'hot spots', that is, very specific subgroup patterns,



Hauptmenü   Name: John Smith   Geb.Dat.: 01.01.1950   Fall speichern

1	()	+	()	()	()	+	+	+	Zahn vorhanden	+	+	+	+	()	+	()	()	1
2		-				-	-	-	Röntgenologische Veränderungen	-	-	-	-		-			2
3		--Em				-ESx	+-	+-	Vitalität, Perkussion, Endo	+-	+-	+-	-ESx		+-			3
4		>5 +				>5 -	>5 +	>5 +	Klinische Krone	>5 +	>3 +	>5 +	>3 o		>3 +			4
5		0				0	0	0	Lockerungsgrad ⓘ	0	0	0	0		0			5
6		<=15				<=15	<=15	<=15	Attachmentloss	<=15	<=15	<=15	<=15					
7		W0W				W	W	W	Wurzelanzahl	W	W	W	W					
8		>2K				>2K	>2K	>2K	Wurzellänge	>2K	>2K	>2K	>2K					
9		+				+	+	+	Wurzelkaries	+	+	+	+					
10									Angulation									
11									Elongation / Extrusion									11
12									Zahnstellung im Zahnbogen									12
13		10.0				135.0	15.0	15.0	Zahn-Score	15.0	20.0	15.0	135.0		10.0			13
14		+				x	+	+	Zahnbewertung	+	+	+	x		+			14
15		15.0				1035.0	15.0	15.0	Pfeiler-Score	15.0	20.0	15.0	1035.0		10.0			15
16		P +				P -	P +	P +	Pfeilereignung	P +	P ?	P +	P -		P +			16
	18	17	16	15	14	13	12	11		21	22	23	24	25	26	27	28	

● Grad 0
○ Grad I
○ Grad II
○ Grad III

Fig. 3 Part of a screenshot of the knowledge-based dialog of the DENTISCONSULT system (in German), showing the documented findings for the upper jaw. The questions are arranged in a questionnaire that is similar to a conventional dental documentation sheet. Then, the examiner can conveniently enter the documented findings using drop-down menus

Table 1 Examples of discovered subgroups: The shown factor combinations significantly increased the incorrect derivation of the diagnosis 'EX' (tooth extraction)

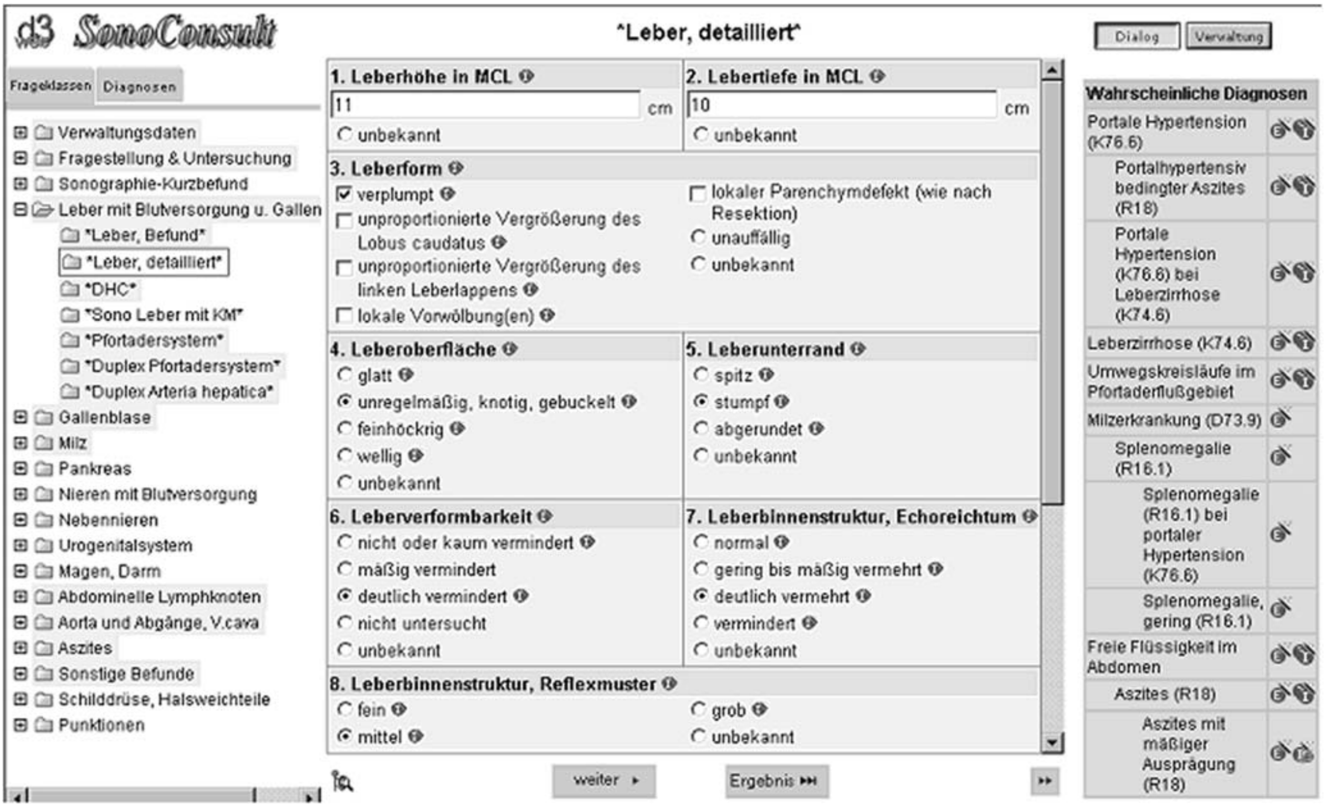
No.	Subgroup description	Diagnosis
1	abnormal x-ray = only apical	EX
2	tooth lax = medium ∧ root length = longer than crown length	EX
3	root caries = minor or on surface	EX
4	tooth lax = none ∧ attachmentloss = strong ∧ endodontic state = possible	EX

needed to be analyzed in detail, either statistically or by viewing the detailed cases, in order to decide about the correct adaptations with respect to the relations modeled in the knowledge base. Examples for such quite small subgroups are given by the subgroup descriptions #3 and #4. Subgroup #3 is similar to subgroup #1 while it was observed in a significantly smaller number of cases. The highly specific subgroup #4 is an exception similar to subgroup #2: Factor *tooth lax* = *none* observed in that situation is a strong factor negative for extraction. Since these factors were observed in very small sub-populations, the presented analysis techniques proved highly useful in determining and especially validating such relations: Manual inspection of the prototypical pattern cases was a key feature for the domain specialist performing the analysis. Figure 2 in Sect. 3.3 shows an example of such a case.

The method allowed for a comprehensive overview on the sub-population defined by a small set of exemplary cases. Furthermore, also the 'drill-down' techniques from factors to sets of cases and for navigating the neighborhood of the a set of retrieved cases proved very helpful during the application. This provided for an easier analysis of the important factor combinations, their contributions and the specific contexts they occurred in.

4.2 Characterizing subgroup patterns in the context of knowledge discovery

For the second application context, cases acquired using the SONOCONSULT [14] system were utilized. SONOCONSULT is a medical documentation and consultation system for sonography which has been developed with the knowledge system D3 [18]. SONOCONSULT is in routine use in



Leber, detailliert

1. Leberhöhe in MCL: 11 cm
○ unbekannt

2. Lebertiefe in MCL: 10 cm
○ unbekannt

3. Leberform
☑ verplumpt
☐ unproportionierte Vergrößerung des Lobus caudatus
☐ unproportionierte Vergrößerung des linken Leberlappens
☐ lokale Vorwölbung(en)
☐ lokaler Parenchymdefekt (wie nach Resektion)
○ unauffällig
○ unbekannt

4. Leberoberfläche
○ glatt
● unregelmäßig, knotig, gebuckelt
○ feinhöckrig
○ wellig
○ unbekannt

5. Leberunterrand
○ spitz
● stumpf
○ abgerundet
○ unbekannt

6. Leberverformbarkeit
○ nicht oder kaum vermindert
○ mäßig vermindert
● deutlich vermindert
○ nicht untersucht
○ unbekannt

7. Leberbinnenstruktur, Echoreichtum
○ normal
○ gering bis mäßig vermehrt
● deutlich vermehrt
○ vermindert
○ unbekannt

8. Leberbinnenstruktur, Reflexmuster
○ fein
● mittel
○ grob
○ unbekannt

Wahrscheinliche Diagnosen
Portale Hypertension (K76.6)
Portalhypertensiv bedingter Aszites (R18)
Portale Hypertension (K76.6) bei Leberzirrhose (K74.6)
Leberzirrhose (K74.6)
Umwegskreisläufe im Pfortaderflußgebiet
Milzerkrankung (D73.9)
Splenomegalie (R16.1)
Splenomegalie (R16.1) bei portaler Hypertension (K76.6)
Splenomegalie, gering (R16.1)
Freie Flüssigkeit im Abdomen
Aszites (R18)
Aszites mit mäßiger Ausprägung (R18)

Fig. 4 Structured data-acquisition using SonoConsult (screenshot in German). The middle part shows a questionnaire for detailed documentation of the liver: It is dynamic, that is, depending on the answers of previous questions, additional questions may be asked. The *left part* enables the user to select questionnaires for the various organs. The *right part* shows diagnoses that are automatically inferred from the data

the DRK-hospital in Berlin/Köpenick and in the Würzburg University Hospital. The documented cases contain detailed descriptions of findings of the examination(s), together with the inferred diagnoses, and additional meta-information. The derived diagnoses of a case are usually correct with respect to the documented findings as shown in a medical evaluation, cf. [14], resulting in a high-quality case base with detailed case descriptions. Figure 4 shows a screenshot of the dialog of the SonoConsult system being used for structured data acquisition.

Currently, the collected SonoConsult case base consists of about 11000 cases. Due to the structured data gathering strategy and the high quality of the case descriptions the system and the collected case base provide excellent opportunities for data analysis and knowledge discovery.

Parts of the collected case base of SonoConsult were already utilized for knowledge discovery and data analysis using subgroup mining methods [6, 7]. The methods were applied in order to discover interesting clinical relations between different organ systems since the intra-organ relations are usually known in the domain of sonography. Furthermore, subgroup mining was applied for quality control with respect to the documentation habits of the sonographic examiners. Then, novel relations between different organ systems could be discovered and documentation profiles for certain examiners could be obtained. Both the relations and the profiles are represented by interesting subgroup patterns. However, the demand for a deeper inspection and characterization of the discovered subgroup patterns in terms of real cases and the further need for identifying related meta-information contained in the cases motivated the development of the presented techniques. The analysis and presentation of the subgroup characteristics in the form of prototypical pattern cases enables an intuitive summary for the clinicians, and furthermore decreases the workload when inspecting the set of cases. Since not all cases contained in the target space of the subgroup need to be analyzed, the analysis effort, that is, the time spent for the analysis can be significantly decreased. This usually also increases the acceptance rate in clinical applications.

In the described clinical context, the proposed methods for characterization and analysis of subgroup patterns provide several opportunities for analysis: The medical experts can directly locate interesting contexts, that is, exemplary cases of specific patients, and typical case descriptions for a specific subgroup pattern. The generated prototypical cases are then applied in order to obtain a summary of the typical problem setting of a subgroup pattern, and for subsequently identifying relevant meta-information contained in the characteristic set of cases. In this context, the users can easily discover relevant meta-information, for example, certain examiners and images associated with a given subgroup pattern.

5 Conclusion

This paper introduces a case-based method for subgroup analysis and characterization. The approach first characterizes a subgroup in terms of its characteristic factors, ranks them, retrieves corresponding typical or extreme cases and finally combines them into a prototypical pattern case. During the merge step, background knowledge can be applied for further improving the interestingness of the prototypical cases, depending on the requirements of the user. Using this representation the user can obtain a comprehensive overview of the problem setting of the subgroup pattern. Furthermore, using 'drill-down' operations on the set of cases, interesting meta-information contained in the characteristic (real) cases can be identified.

In the future, the authors plan to investigate further techniques for subgroup characterization and summarization, for example, based on clustering techniques and other condensed forms of sets of subgroups.

Acknowledgements The authors want to thank Achim Hemsing and Prof. Ernst-Jürgen Richter from the department of prosthodontics at the Würzburg University Hospital, Prof. Hans-Peter Buscher from the department of internal medicine at the DRK-Klinik Berlin-Köpenick, and Hardi Lührs from the department of sonography at the Würzburg University Hospital for their medical expertise and analysis while performing the case studies of this research project. This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-1.

References

1. Aamodt A, Plaza E (1994) Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun 7(1):39–59
2. Aha DW (1992) Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. Int J Man–Mach Stud 36(2):267–287
3. Atzmueller M, Baumeister J, Hemsing A, Richter E-J, Puppe F (2005) Subgroup mining for interactive knowledge refinement. In: Proceedings of the 10th conference on artificial intelligence in medicine (AIME 05). Lecture notes in artificial intelligence, vol 3581. Springer, Berlin, pp 453–462
4. Atzmueller M, Baumeister J, Puppe F (2006) Introspective subgroup analysis for interactive knowledge refinement. In: Sutcliffe G, Goebel R (eds) Proceedings of the 19th international Florida artificial intelligence research society conference 2006 (FLAIRS-2006). AAAI, Menlo Park, pp 402–407
5. Atzmueller M, Baumeister J, Puppe F (2006) Semi-automatic learning of simple diagnostic scores utilizing complexity measures. Artif Intell Med 37(1):19–30, Special issue on intelligent data analysis in medicine
6. Atzmueller M, Puppe F, Buscher H-P (2005) Exploiting background knowledge for knowledge-intensive subgroup discovery. In: Proceedings of the 19th international joint conference on artificial intelligence (IJCAI-05), Edinburgh, Scotland, pp 647–652
7. Atzmueller M, Puppe F, Buscher H-P (2005) Profiling examiners using intelligent subgroup mining. In: Proceedings of the 10th international workshop on intelligent data analysis in medicine and pharmacology (IDAMAP-2005), Aberdeen, Scotland, pp 46–51
8. Bareiss R (1989) Exemplar-based knowledge acquisition: a unified approach to concept representation, classification, and learning. Academic, San Diego
9. Bartsch-Spörl B, Lenz M, Hübner A (1999) Case-based reasoning: survey and future directions. In: XPS-99: knowledge-based systems—survey and future directions, proceedings of the 5th biannual German conference on knowledge-based systems, pp 67–89
10. Bergmann R (2002) Experience management: foundations, development methodology, and Internet-based applications. Springer, Berlin
11. Gamberger D, Krstacic A, Krstacic G, Lavrac N, Sebag M (2005) Data analysis based on subgroup discovery: experiments in brain ischaemia domain. In: Proceedings of the 10th international workshop on intelligent data analysis in medicine and pharmacology (IDAMAP-2005), Aberdeen, Scotland, pp 52–56
12. Gamberger D, Lavrac N, Krstacic G (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artif Intell Med 28:27–57
13. Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning: In: Proceedings of the 17th international conference on machine learning. Kaufmann, San Francisco, pp 359–366
14. Huettig M, Buscher G, Menzel T, Scheppach W, Puppe F, Buscher H-P (2004) A diagnostic expert system for structured reports, quality assessment, and training of residents in sonography. Med Klin 99(3):117–122
15. Klösgen W (1996) Explora: A multipattern and multistrategy discovery assistant. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining, AAAI, Menlo Park, pp 249–271
16. Lavrac N, Gamberger D, Flach P (2002) Subgroup discovery for actionable knowledge generation: shortcomings of classification rule learning and the lessons learned. In: Lavrac N, Motoda H, Fawcett T (eds) Proceedings of the ICML 2002 workshop on data mining: lessons learned, July 2002
17. McSherry D (2002) Diversity-conscious retrieval. In: Proceedings 6th European conference on advances in case-based reasoning. Springer, Berlin, pp 219–233
18. Puppe F (1998) Knowledge reuse among diagnostic problem-solving methods in the shell-kit D3. Int J Human–Comput Stud 49:627–649
19. Schmidt R, Gierl L (2001) Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes. Artif Intell Med 23(2):171–186
20. Smyth B, McClave P (2001) Similarity vs. diversity. In: Proceedings of the 4th International conference on case-based reasoning (ICCBR 01). Springer, Berlin, pp 347–361
21. Witten IH, Frank E (1999) Data mining: practical machine learning tools and techniques with Java implementations. Kaufmann, Los Altos
22. Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Komorowski J, Zytkow J (eds) Proceedings of the 1st European symposium on principles of data mining and knowledge discovery (PKDD-97). Springer, Berlin, pp 78–87